

**SORT MIDDLE, SCREEN SPACE, GRAPHICS GEOMETRY  
COMPRESSION THROUGH REDUNDANCY ELIMINATION**

**CROSS REFERENCE TO RELATED APPLICATION**

5           This is a continuation application of co-pending application number  
09/412,898 filed on October 5, 1999, which is hereby incorporated by reference  
herein.

**BACKGROUND OF THE INVENTION**

1. Field of Invention

10           The present invention relates generally to the field of computer graphics and  
pertains more particularly to sort middle, screen space, graphics geometry  
compression through redundancy elimination.

2. Discussion of the Prior Art

Modern computer systems have become increasingly graphics intensive.

15   Dedicated special purpose memories and hardware have been developed to meet this  
need. The majority of graphics systems are built to accelerate the standard graphics  
pipeline. As an example FIG. 1 shows a block diagram of a graphics pipeline 10  
having geometry processing 12 and rasterization 14. The pipeline 10 processes a three  
dimensional (3D) database 16 having geometric objects 18 and textures 20 into a  
20   display image 22 which may be shown to the user on a display monitor (not shown).  
The geometric objects 18 are typically polygons but may also be triangles, lines, and  
points. These are also known as primitives. For a particular primitive, its vertices or  
geometrical point locations ( $p_x$ ,  $p_y$ ,  $p_z$ ) are typically IEEE floating point values. Also,  
attributes such as colors, normals, and texture coordinates may be specified at each

-2-

point or for an entire primitive. In binary, the floating point value is made up of one bit for the sign (S), eight bits for the exponent (E or exp), and twenty-three bits for the significand for a total of thirty-two bits. In the case of a triangle, there are three vertices generally having horizontal (x), vertical (y), and depth (z) coordinates in the form of  $(v_{0x}, v_{0y}, v_{0z})$ ,  $(v_{1x}, v_{1y}, v_{1z})$ , and  $(v_{2x}, v_{2y}, v_{2z})$ . These primitive vertex data of the geometric objects are sent for geometry processing 12.

During geometry processing 12, this primitive vertex data undergoes modeling transforms, viewing transforms, and viewpoint transforms. The resulting screen space primitives are specified by a number of setup variables. Now one has  $(v_{0x}, v_{0y}, v_{0z})$  in screen space. Setup variables include the starting values and slope of all of the parameters of the primitive. The number, precision, and type of setup variables will vary from application to application. Geometry processing output may include the screen space locations (x and y) of the triangle vertices, the shaded color values (red, green, blue, alpha), the depths (z), the slope of the edges, colors, opacities, and depths across scanlines, triangles, or both. Calculation of the setup variables may be done at the geometry processing stage or at the rasterization stage. This geometry information varies by the rasterization process used and the techniques disclosed below can work with many varieties of rasterization.

Turning now to FIG. 2, a diagram of scan conversion of two triangle types is shown. FIG. 2 is for illustration purposes only as there exist many equivalent methods to rasterize triangles. For both triangle types, edge 1, the longest side, is the primary vertical edge on which scan conversion starts. All scan conversion proceeds in horizontal lines across the triangle as shown. Textures 20 from FIG. 1 may be used as an auxiliary method for looking up a color of a pixel during the scan conversion

process. The texture is literally painted onto the polygon by performing an inverse transform lookup from the pixel coordinate, to the polygon coordinate, into the texture space. Finally, the resulting rasterized data is displayed as the display image 22.

In general, the efficient utilization of the special purpose memories and  
5 hardware of the graphics system has not been fully considered. As graphics technologies and products become more prevalent in consumer areas, slim profit margins are demanding the utmost efficiencies. Images rendered from a large number of triangles can create a fundamental bottleneck in the system. Approaches to relieve this situation are to either modify the geometry to create fewer triangles or to  
10 compress the triangles that are given. While compression has been used to make many technologies possible, such as communications and video technologies, compression has not been used in conventional graphics systems. Compression is necessary to achieve efficiency in database representation to conserve disk storage, system memory, cache memory, and special purpose memories. Additional savings  
15 can be achieved by conserving system bus bandwidths, which becomes more critical with unified computer architectures where graphics and system memory are one and the same.

A definite need exists for a system having an ability to meet the efficiency requirements of graphics intensive computer systems. In particular, a need exists for a  
20 system which is capable of compressing geometric data in a skillful manner. Ideally, such a system would have a lower cost and a higher productivity than conventional systems. With a system of this type, system performance can be enhanced. A primary purpose of the present invention is to solve this need and provide further, related advantages.

## SUMMARY OF THE INVENTION

A geometry compression method is disclosed for sort middle, screen space, graphics of the standard graphics pipeline. The pipeline processes a 3D database  
5 having geometric objects such as triangles and textures into a display image which may be shown to the user on a display monitor. Lossless compression is achieved through redundancy elimination. Triangles are processed following their transformation to screen space, so that the vertex world 3D locations are determined in their projection to the screen 2D locations. Triangles may also be processed by  
10 back projecting the screen space scanlines to test locations against the world space triangles. The general technique is to identify the portions of the data that have little or no effect on the rendered output and remove them during compression. Specific examples disclosed include full packing, constant color, delta coding, edge sharing, slope coding, and color quantization.

15

## BRIEF DESCRIPTION OF THE DRAWING

The above and other objects and advantages of the present invention will be more readily appreciated from the following detailed description when read in conjunction with the accompanying drawing, wherein:

20 FIG. 1 is a block diagram of a prior art graphics pipeline having geometry processing and rasterization;

FIG. 2 is a diagram of a known scan conversion of two triangle types;

FIG. 3 is a block diagram of the graphics pipeline of the type shown in FIG. 1 having geometry compression and geometry decompression according to one embodiment of the present invention;

FIG. 4 is a diagram of two triangles having a shared edge;

5 FIG. 5 is a diagram illustrating four reference heights and widths of a triangle;

FIG. 6 is a flow diagram of setup for a color quantization graphics compression method according to one embodiment of the present invention; and

FIG. 7 is a flow diagram of a color quantization graphics compression method according to one embodiment of the present invention.

10

#### **DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS**

Turning now to FIG. 3, a block diagram of the graphics pipeline 10 according to FIG. 1 having geometry compression 24 and geometry decompression 25 according to one embodiment of the present invention is shown. The geometry compression 24 is shown situated between the geometry processing 12 and the rasterization 14 and operates on the setup variables for the screen space primitives of the sort middle traffic as described above. The geometry decompression 25 follows the geometry compression 24 and precedes the rasterization 14. Alternatively, the geometry compression 24 and geometry decompression 25 may be partially or wholly integrated into the geometry processing 12, the rasterization 14, or both. Among the advantages of locating the geometry compression 24 and geometry decompression 25 at this point in the graphics pipeline 10 is that no changes to the application program interface (API) are required thus making the present invention independent of which API is used. Further, more precise knowledge can be utilized from the screen space than the

world space. Further still, this location corresponds to a potential bottleneck point as there is an expansion of data from this point forward in the graphics pipeline 10. The geometry compression 24 may perform either or both of the two types of compression which are lossless and lossy. Lossless compression refers to compression with no effect on the rendered output. Unused and redundant data are prime candidates for lossless compression by selective elimination of this data. Lossy compression refers to compression with some effect on the rendered output. Generally the effect will be to such a small extent as to be acceptable on balance. What one deems to be acceptable will depend on the situation.

10           In this application, lossless compression refers to encoding of the setup variables in such a way that guarantees that the rendered output generated from the encoded data is bit-wise identical to that generated from the original. Thus portions of the data that are unused or redundant and therefore have no effect on the rendered output can be removed by the geometry compression 24.

15           Recall from above that the vertices of a primitive are typically floating point values made up of thirty-two bits. Under a typical graphics pipeline protocol, there are  $n$  bits, indexed from zero to  $n-1$ , in a word containing the values. The setup variables may be represented as floating point or fixed point with varying precision as determined by need. For example,  $z$  may be represented as forty bits, while color components may be represented as twenty-four bits. In the present embodiment of the invention, setup variables are considered to be fixed point values with differing precision requirements. Assume for example that an elementary protocol for communicating setup data is to transmit the relevant parameters/variables as a sequence of words along with a few bits of overhead per word to identify the type of

-7-

parameter being conveyed. Part of this overhead is known as the type field and is used to specify the type of setup variable. Recall also from above that the type of setup variables will vary from application to application. The type field permits maximum flexibility in ordering the transmission of parameters as well as in selecting the subset of parameters necessary for describing a primitive.. However, this flexibility may not be necessary in many applications. For example, this flexibility is not necessary when the system only has a small number of primitives and a small number of rendering modes.

A geometry compression method that will be referred to as full packing can be utilized by a system when the type field is not necessary. The method includes defining a packet type corresponding to a setup variable type, identifying the type of a set of setup variables, and bundling the setup variables into a packet of the corresponding packet type. Each packet includes a header and a set of setup variables in a predefined order. In this way, the type field for each individual setup variable is eliminated. The actual savings will depend on the situation, but the number of additional bits of the header is substantially less than the number of bits eliminated for all of the type fields for a set of setup variables. Under full packing, over 100 bits would be eliminated in exchange for the header bits. For an image containing hundreds or thousands of primitives, the total savings would be substantial.

Among the setup variables generated by geometry processing is a group of six slope values that specify two slope values for each of the primary colors red (R), green (G), and blue (B). The slope values are added to the starting value to compute the color of each pixel falling inside the triangle. The slopes roughly specify the change in each color per horizontal and vertical pixel distance. However, these slope values



may not be necessary in some applications. For example, these slope values are not necessary when the triangle exhibits flat shading as indicated by zero values for all six slope values.

5 A geometry compression method that will be referred to as constant color can be utilized by a system when these slope values are not necessary. The method includes identifying when all six of the slope values have a zero value and utilizing a bit to indicate that such a condition exists. In this way, the six slope values can be eliminated for the triangle in question. Under constant color, hundreds of bits will be eliminated in exchange for one bit. For an image containing hundreds or thousands of  
10 constant color primitives, the total savings would be substantial.

Also among the setup variables generated by geometry processing is a group of six values that specify the starting x, the starting y, and the stopping y coordinates of the three edges of the triangle. However, the information conveyed by these coordinate values may be expressed in a different manner in some applications. For  
15 example, when the triangles are predominantly small in size, the coordinate values could be expressed in terms of an anchor coordinate value and two differences relative to the anchor coordinate value.

A geometry compression method that will be referred to as delta coding can be utilized by a system when the triangles are predominantly small in size. The method  
20 includes selecting an anchor coordinate value in each of the x and y directions, calculating the differences between the anchor coordinate and the other two coordinates in each of the x and y directions, and encoding the differences. After computing the four differences, these differences are encoded using an adaptive Golomb coding technique which is well known in the art. Other encoding techniques



are also possible. In this way, the six coordinate values can be compressed for the triangle in question. Under delta coding, over 100 bits will be eliminated in exchange for four compressed differences. For an image containing hundreds or thousands of primitives, the total savings would be substantial.

5           Turning now to FIG. 4, a diagram of two triangles having a shared edge 26 is shown. Those of ordinary skill in the art will realize that consecutive triangles of an image often share a common edge. Conventionally, each of the two triangles would have all of their respective setup variables transmitted. It is possible however to derive certain of the setup variables for the second triangle from those of the first  
10   triangle. Which setup variables can be derived will depend on which edges are shared. There are nine possible matches.

A geometry compression method that will be referred to as edge sharing can be utilized by a system when two triangles have a shared edge. The method includes identifying when two triangles have a shared edge and utilizing four bits to indicate  
15   that such a condition exists and which edge is shared. In the event of a shared edge, two y coordinates, one x coordinate, and one edge slope of the second triangle can be derived from the corresponding setup variables for the shared edge of the first triangle. In this way, the corresponding five setup variables can be eliminated for the second triangle. Under edge sharing, hundreds of bits will be eliminated in exchange for four  
20   bits. For an image containing hundreds or thousands of shared edge primitives, the total savings would be substantial.

Also among the setup variables generated by geometry processing is a group of three values that specify the edge slopes of the triangle. However, the information conveyed by these edge slopes may be expressed with more precision than necessary

-10-

in some applications. For example, when the integer portion of the absolute value of the slope is zero, several bits of integer precision are wasted.

A geometry compression method that will be referred to as slope coding can be utilized by a system when the precision of the edge slopes is not necessary. The method includes identifying when the absolute value of a slope is larger than a  
5 predetermined threshold such as two and transmitting the integer portion of the slope using full precision when such a condition exists. In either case, one bit is used to indicate which condition exists. In this way, one bit is added to the overhead for the edge slope in the case of a large slope and several bits can be eliminated for the case  
10 of a small slope. Under slope coding, as many as thirty-six bits may be eliminated in exchange for three bits. For an image containing hundreds or thousands of small edge slope primitives, the total savings would be substantial.

Also among the setup variables generated by geometry processing is a group of nine values that specify the three color start values and two slope values for each of  
15 the primary colors red (R), green (G), and blue (B). The slope values are added to the starting value to compute the color of each pixel falling inside the triangle. The slopes roughly specify the change in each color per horizontal and vertical pixel distance. However, these start and slope values may have a precision that is greater than necessary in some applications to prevent errors from accumulating to greater than  
20 one gray level in the rendered output. For example, these start and slope values are not necessary when the triangle is small in size, that is, many times smaller than the visible display area.

A geometry compression method that will be referred to as color quantization can be utilized by a system when these start and slope values are more precise than

necessary. The method includes identifying when a triangle is substantially smaller than the visible display area and quantizing the start and slope setup variables to integer values such that the precision varies with the primitive size and guarantees a maximum error of a predetermined portion of a gray level throughout each rasterized primitive. For example, the gray level error may be chosen to be 1/4, 1/8, 1/16, etc. In this way, the precision of the nine start and slope values can be tailored for the triangle in question. It should be noted that a 1/4 gray level error may introduce distortion into the rendered output relative to the uncompressed rendered output. In practice, it was found that this error was sufficiently small so as to be imperceptible in the final output.

The key to tailoring the precision of the nine start and slope values for the triangle in question is to be able to determine the final error and the number of times the slopes are accumulated. This is strongly related to the size of the triangle. The rasterizers compute the color of any pixel according to the formula

$$c = c_0 + N_e c_e + N_x c_x, \quad (1)$$

where  $c_0$  is the starting color value,  $c_e$  is the color edge slope,  $c_x$  is the color x scanline slope, and  $N_e$  and  $N_x$  are respectively the number of times that  $c_e$  and  $c_x$  are accumulated to obtain  $c$ . If the parameters  $c_0$ ,  $c_e$ , and  $c_x$  are approximated by  $\hat{c}_0$ ,  $\hat{c}_e$ , and  $\hat{c}_x$ , then the error  $\varepsilon$  in the final color value can be given an upper bound by the formula

$$\varepsilon \leq \varepsilon_0 + N_e \varepsilon_e + N_x \varepsilon_x, \quad (2)$$

where  $\varepsilon_0 = |c_0 - \hat{c}_0|$ ,  $\varepsilon_e = |c_e - \hat{c}_e|$ , and  $\varepsilon_x = |c_x - \hat{c}_x|$  are the errors in the approximation of the parameters. If  $R_0$ ,  $R_e$ , and  $R_x$  are the fractional bits used to approximate  $c_0$ ,  $c_e$ , and  $c_x$ , respectively, then the final error is guaranteed to be smaller than  $\varepsilon$  if

-12-

$$\varepsilon \geq 2^{-(R_0+1)} + N_e 2^{-(R_e+1)} + N_x 2^{-(R_x+1)} \quad (3)$$

It can be shown that the minimum of  $R_0+R_e+R_x$  subject to the constraint of EQ. (3) is achieved by making

$$R_0 = -\log_2(\varepsilon/3) - 1, \quad (4)$$

$$5 \quad R_e = -\log_2(\varepsilon/3N_e) - 1, \quad (5)$$

and

$$R_x = -\log_2(\varepsilon/3N_x) - 1. \quad (6)$$

For simplification,  $R_e$  can be set equal to  $R_x$ . As a result, the fractional bits become

$$R_0 = -\log_2(\varepsilon/2) - 1, \quad (7)$$

$$10 \quad R_e = R_x = -\log_2(\varepsilon/(2(N_e + N_x))) - 1. \quad (8)$$

The error  $\varepsilon$  can be selected to have a value of 1/4, 1/8, 1/16, 1/32, etc. Acceptable results were obtained using the value of 1/4.

Turning now to FIG. 5, a diagram illustrating four reference heights and widths of a triangle is shown. Each length is measured in pixel center distances. The length  $H$  is the absolute height of the triangle. The length  $H'$  is the height of the triangle from its widest point. The length  $W$  is the width of the triangle at its widest point. The length  $W'$  is the absolute width of the triangle. Using these four lengths, two bounds on the sum of  $N_e + N_x$  can be determined. A first bound is given by the formula

$$20 \quad N_e + N_x \leq \max(2H, 2H' + W). \quad (9)$$

A second simpler but weaker bound is given by the formula

$$N_e + N_x \leq (2H + W). \quad (10)$$

Turning now to FIG. 6, a flow diagram of setup for a color quantization graphics compression method according to one embodiment of the present invention

is shown. The setup is preferably done one time as a preprocess at the time of the implementation of the hardware or firmware. The process begins at START. At block 28, the process selects an error bound  $\epsilon$  for the compression. At block 30, the process calculates the number of bits needed for the starting color value (see EQ. 7).  
5 These preselected and precalculated values will impact the method of FIG. 7 which is described next.

Turning now to FIG. 7, a flow diagram of a color quantization graphics compression method according to one embodiment of the present invention is shown.

The process begins at START for each triangle and returns to this point until there are  
10 no more triangles. The same precision may be used for all attributes of the same start precision, such as red, green, blue, and alpha. At block 32, the process computes the heights and widths of the triangle. Recall from FIG. 5 above that these include the absolute height of the triangle, the height of the triangle from its widest point, the width of the triangle at its widest point, and the absolute width of the triangle.  
15 However, depending on the bound used in the next block of the process, all four of these may not be needed. At block 34, the process estimates the bound on the sum of  $N_e + N_x$  (see EQ. 9 or 10). At block 36, the process calculates the number of bits needed for the edge slope value and the x scanline slope value (see EQ. 8). Finally, at  
20 block 38, the process encodes the starting color value, the edge slope value, and the x scanline slope value using the calculated number of bits. Under color quantization, unnecessary precision bits can be eliminated. More bits are typically used for z and fewer bits for color. Possible candidate setup variables for color quantization include color (red, green, and blue), opacity (alpha), depth (z), texture variables, and normals.

For an image containing hundreds or thousands of small primitives, the total savings would be substantial.

The specific examples of geometry compression disclosed above included full packing, constant color, delta coding, edge sharing, slope coding, and color

5 quantization. Taken individually, each represents a substantial savings. It should also be noted that the techniques can be combined for even greater savings overall. Taken in combination, compression ratios as high as two to one have been achieved in practice. This represents far more efficient utilization of the special purpose memories and hardware of the graphics system.

10 While the invention has been illustrated and described by means of specific embodiments, it is to be understood that numerous changes and modifications may be made therein without departing from the spirit and scope of the invention as defined in the appended claims and equivalents thereof.